Internship Proposal

Design optimization of AI/ML Hardware Accelerators

Hosting Institution: ***TIMA Laboratory***
Supervisor: ***Mounir BENABDENBI and Olivier MULLER***
Duration of contract: ***6 months***
Contact: Mounir BENABDENBI (Mounir.Benabdenbi@univ-grenoble-alpes.fr)

Context:

During the last decades, the semiconductor industry showed many improvements in terms of energy efficiency of produced devices: multi-core designs, (ultra) low power components, etc. However, the energy consumption of computer systems is still growing at an alarming rate. In addition, a large number of applications usually referred as Recognition, Mining and Synthesis (RMS), have emerged and are gaining more and more popularity. From mobile and Internet of Things (IoT) to large-scale data centers, they account now for a significant and growing portion of global computational resources and energy consumption.

It is then mandatory to break the power wall by improving the energy efficiency of silicon devices. This is actually more critical since semiconductor technology continue to scale down to nanometer transistors: better performances but at the price of an increasing energy inefficiency.

Hopefully, many applications, such as the RMS ones, often feature intrinsic error-resilience properties. Indeed, these applications do not have to provide a unique or exact result. For example, in video applications, some types of errors such as locally adaptive quantization can be easily tolerated, as long as the error remains below visual threshold of human visual perception.

Based on these observations, in the last decade, a very promising solution known as "Approximate Computing" (AxC), has gained more and more interest in the scientific community both in the industry and in academia. AxC is based on the intuitive observation that, while performing exact computation requires a high amount of resources, allowing selective approximation or occasional violation of the specifications can provide gains in power consumption, more than one magnitude order. Or, for the same amount of consumption, performances can be enhanced. Various applications of AxC were surveyed such as data analytics, scientific computing, multimedia, signal processing, machine learning and so forth.

Looking at the state of the art [1], the proposed AxC techniques can be classified as: AxC software (e.g., reduced algorithm iterations, OS task management), AxC architecture (e.g., processors with approximate arithmetic blocs) or AxC circuits (e.g., reduced supply voltage, imprecise logic).

Recent works have shown that when looking at AI algorithms, we can consider that they are already naturally approximate. Indeed, data types are often over designed, and simplified coding (BNNs and TNNs) can lead to less complex hardware networks, with a smaller footprint, better performance and reduced power consumption, with a very small impact on the outputs' accuracy.

For CNNs, it was also shown that some layers don't need a maximum precision to deliver an exact output. The challenge is then to finely tune the precision of each layer and/or each arithmetic operator by choosing the right type and size of variables (float, int, or arbitrary non-standard bit-widths). The difficulty is to find the right architectural parameters targeting a Quality Of Results (QoR) for a performance level. This can be done in a first step by a series of trial and error iterations allowing to converge towards a tryptic (performance, QoR, power consumption).

Ideally it may also be done during the training phase leading to a one-shot final architecture with ideal parameters.

Internship Proposal

The purpose of this project is to propose a methodology to generate the best architectural parameters (optimized set of data types) to design energy efficient AI hardware accelerators. The case study will target the implementation of an AI application based on CNNs.

For that, we will implement the first method (trial-error) using an architectural exploration and evaluation tool developed at TIMA [2]. This tool is based on a description of the hardware architecture in the object-oriented Chisel/Scala language. Thanks to Chisel, we expect a simplified and automated generation (and evaluation) of different versions of the architecture, each with different custom data types and formats.

Forecasted internship Tasks:

The work of the internship will consist in accomplishing the following tasks:

- Task 1: Selection and study of the CNN-based application
- Task 2: Using the software version of the application, evaluation of the application outputs' accuracy loss for different data types at different layers of the CNN
- Task 3: Design of the hardware implementation using the CHISEL language
- Task 4: Generation and evaluation of different versions of the hardware accelerator using the TIMA's tool. The synthesis will target an implementation on FPGAs.
- Task 5: Through the knowledge acquired from the software and hardware experiments propose a methodology/algorithm to select the best architectural parameters for the set (power consumption, performance, accuracy loss)

We are looking for a highly-motivated Engineering School or M2 Masters student. Applicants must have a Master 1 (or equivalent) in a related field of microelectronics. Interpersonal skills, dynamism, rigor and teamwork abilities will be appreciated. Candidates should be fluent in English and/or in French and have good English writing skills. Our team welcomes applicants with diverse backgrounds and experiences. We regard gender equality and diversity as strength and an asset.

Required skills: Strong knowledge on programming languages, hardware design and simulation (FPGA). AI/ML knowledge is greatly appreciated as an experience with the Chisel language, but they are not mandatory.

For highly motivated student, with a good CV, the internship may be followed by a PhD grant application.

About TIMA:

TIMA Laboratory is a public joint research laboratory located in Grenoble, France, and held jointly by Institut Polytechnique de Grenoble (Grenoble INP), University Grenoble-Alpes and French National Research Council (CNRS).

Application:

Please send your CV and letter of motivation, together with references to:

Mounir Benabdenbi (Mounir.Benabdenbi@univ-grenoble-alpes.fr)

[1] Anghel L., Benabdenbi M., Bosio A., Traiola M., Vatajelu I., "Test and Reliability in Approximate Computing" Journal of Electronic Testing: Theory and Applications, Ed. Springer, Vol. 34, No. 4, pp. 375-387, DOI: 10.1007/s10836-018-5734-9, août 2018

[2] Bruno Ferres, "Leveraging Hardware Construction Languages for Flexible Design Space Exploration on FPGA," PhD Dissertation, 2022, https://hal.archives-ouvertes.fr/tel-03709710